

# Spatio-temporal spike trains analysis for large scale networks using maximum entropy principle and Monte-Carlo method

Hassan Nasser <sup>\*</sup>, Olivier Marre <sup>†</sup>, Bruno Cessac <sup>‡</sup>

September 19, 2012

## Abstract

Understanding the dynamics of neural networks is a major challenge in experimental neuroscience. For that purpose, a modelling of the recorded activity that reproduces the main statistics of the data is required. In a first part, we present a review on recent results dealing with spike train statistics analysis using maximum entropy models (MaxEnt). Most of these studies have been focusing on modelling synchronous spike patterns, leaving aside the temporal dynamics of the neural activity. However, the maximum entropy principle can be generalized to the temporal case, leading to Markovian models where memory effects and time correlations in the dynamics are properly taken into account. In a second part, we present a new method based on Monte-Carlo sampling which is suited for the fitting of large-scale spatio-temporal MaxEnt models. The formalism and the tools presented here will be essential to fit MaxEnt spatio-temporal models to large neural ensembles.

## 1 Introduction

The structure of the cortical activity, and its relevance to sensory stimuli or motor planning, have been subject to long standing debate. While some studies tend to demonstrate that the majority of the information conveyed

---

<sup>\*</sup>NeuroMathComp, INRIA, 2004 Route des Lucioles, 06902 Sophia-Antipolis, France.

<sup>†</sup>Institut de la Vision, 17 rue Moreau, 75012 Paris, France.

<sup>‡</sup>NeuroMathComp, INRIA, 2004 Route des Lucioles, 06902 Sophia-Antipolis, France.

by neurons is contained in the mean firing rate [59], other works have shown evidence for a role of the higher order neural assemblies in neural coding ([65], [77], [1], [33] and [24]).

Many single cell studies have reported an irregular spiking activity which seems to be very close to a Poisson process; concluding that the activity spans a very large state space. Several studies claim that some specific patterns, called “cortical songs”, appear in a recurrent fashion ([27]), but their existence is controversial ([40] and [34]), suggesting that the size of the state space explored by the activity could be smaller than expected. This point requires an accurate description of the neural activity of populations of neurons ([66], [43], [32], [76] and [31]).

These controversies partially originate from the fact that characterizing the statistics of the neural activity observed during the simultaneous recording of several neurons is challenging, since the number of possible patterns grows exponentially with the number of neurons. As a consequence, the probability of each pattern cannot be reliably measured by empirical averaging, and an underlying model is necessary to reduce the number of variable to be estimated. To infer the whole state of the neural network, some attempts have been done to build a hidden dynamical model which would underlie the cortical responses of several recorded neurons. Most of the time, this approach has been used to characterize the activity of neurons during different types of behaviour. Among others, Shenoy and colleagues [60] used a dynamical system to model the activities of multiple neurons recorded in the motor areas. Most of the time, in this approach, the number of neurons largely exceeds the number of parameters. The assumed low dimension of the underlying dynamical system is often due to the low dimension of the behavioural context itself. For example, in a task where a monkey is asked to make a choice between a small number of options (e.g. moving toward one target amongst several), one can expect that the features of the neural activity which are relevant to this task can be described with a number of parameters which is comparable to the number of possible actions.

For more complex tasks or stimuli, the dimension of these models may have to be increased. This would be especially critical in the case of sensory networks stimulated with natural or complex stimuli. For this latter issue, a different strategy has been proposed by Schneidman et al [56] and Shlens et al [62, 63]. Their purpose was to describe the statistics of the retinal activity in response to natural stimuli. They defined a set of values (mean firing rates, correlations...) that must be fitted, and then picked the *least structured* of the models that would satisfy these constraints. This approach, which will be described below, is based on maximum entropy models of the activity. It is interesting to point out that, while the previous approach aims at finding

a useful representation of the activity with the lowest dimension, the maximum entropy approach picks the model with the highest dimension.

In this paper, we first describe the challenge of modelling the statistics of the neural activity, and review the results that were obtained using maximum entropy models. Many studies focused on modelling the synchronous patterns, putting aside the issue of modelling the temporal dynamics of the neural activity. We show why the extent of maximum entropy models to the temporal case raises specific issues, such as treating correctly memory and time correlations, and how they can be solved. The corresponding section (section 2) reviews the maximum entropy approach, and focuses on applying it to general spatio-temporal constraints. We also include a short discussion on other spatio-temporal approaches to spike train statistics such as the Generalized Linear Model [6, 38, 45, 75, 49, 47, 4, 48]. In a second part (section 3) we present a new method, based on Monte-Carlo sampling, which is suited for the fitting of large scale spatio-temporal models. The section 4 provides examples and numerical tests in order to show how far we can go with the Monte-Carlo method and its performance.

## 2 The maximum entropy principle

In this section, we present the maximum entropy principle in a general setting. We first give a set of notations and definitions, then present a brief history of this principle in spike train analysis. Finally, we introduce a framework which allows the handling of general spatio-temporal constraints.

### 2.1 Notations and definitions

#### 2.1.1 Spike trains

We consider a network of  $N$  neurons. We assume there is a minimal time scale  $\delta$ , such that a neuron can fire at most one spike within a time window of size  $\delta$ . To each neuron  $k$  and discrete time  $n$ , we associate a spike variable:  $\omega_k(n) = 1$  if neuron  $k$  fires at time  $n$ , and  $\omega_k(n) = 0$  otherwise. The state of the entire network in time bin  $n$  is thus described by a vector  $\omega(n) \stackrel{\text{def}}{=} [\omega_k(n)]_{k=1}^N$ , called a *spiking pattern*.

A *spike block*, which describes the activity of the whole network between to moment of time  $n_1$  and  $n_2$ , is a finite ordered list of such vectors, written:

$$\omega_{n_1}^{n_2} = \{ \omega(n) \}_{\{n_1 \leq n \leq n_2\}},$$

The *range* of a block is  $n_2 - n_1 + 1$ , the number of time steps from  $n_1$  to  $n_2$ . Here is an example of a spike block of range 5 with  $N = 4$  neurons.

$$\begin{pmatrix} 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

A *spike train* or *raster* is a spike block  $\omega_0^{T-1}$  from some initial time 0 to some final time  $T - 1$ . To simplify notation we simply write  $\omega$  for a spike train. We note  $\Omega = \{0, 1\}^{NT}$  the set of all possible spike trains.

### 2.1.2 Observables

We call *observable* a function:

$$\mathcal{O}(\omega) = \prod_{u=1}^r \omega_{k_u}(n_u), \quad (1)$$

i.e. a product of binary spike events where  $k_u$  is a neuron index and  $n_u$  a time index, with  $u = 1, \dots, r$ , for some integer  $r > 0$ . Typical choices of observables are  $\omega_{k_1}(n_1)$  which is 1 if neuron  $k_1$  fires at time  $n_1$  and which is 0 otherwise;  $\omega_{k_1}(n_1)\omega_{k_2}(n_2)$  which is 1 if neuron  $k_1$  fires at time  $n_1$  and neuron  $k_2$  fires at time  $n_2$  and which is 0 otherwise. Another example is  $\omega_{k_1}(n_1)(1 - \omega_{k_2}(n_2))$  which is 1 if neuron  $k_1$  fires at time  $n_1$  and neuron  $k_2$  is silent at time  $n_2$ . This example emphasizes that observables are able to consider events where some neurons are silent.

We say that an observable  $\mathcal{O}$  has *range*  $R$  if it depends on  $R$  consecutive spike patterns, e.g.  $\mathcal{O}(\omega) = \mathcal{O}(\omega_0^{R-1})$ . We consider here that observables do not depend explicitly on time (*time-translation invariance of observables*). As a consequence, for any time  $n$ ,  $\mathcal{O}(\omega_0^{R-1}) = \mathcal{O}(\omega_n^{n+R-1})$  whenever  $\omega_0^{R-1} = \omega_n^{n+R-1}$ .

### 2.1.3 Spike train statistics

It is common in the study of spike trains to attempt to detect some statistical regularity. Spike trains statistics is assumed to be summarized by a hidden probability  $\mu$  characterizing the probability of *spatio-temporal* spike patterns:  $\mu$  is defined as soon as the probability  $\mu[\omega_{n_1}^{n_2}]$  of any block  $\omega_{n_1}^{n_2}$  is known. We assume that  $\mu$  is time-translation invariant: for any time  $n$ ,  $\mu[\omega_0^{R-1}] = \mu[\omega_n^{n+R-1}]$ , whenever  $\omega_0^{R-1} = \omega_n^{n+R-1}$ .

Equivalently,  $\mu$  allows the computation of the average of the observables. We note  $\mu[\mathcal{O}]$  the average of the observable  $\mathcal{O}$  under  $\mu$ . If  $\mathcal{O}(\omega) = \omega_{k_1}(n_1)$  then  $\mu[\mathcal{O}]$  is the firing rate of neuron  $k_1$  (it does not depend on  $n_1$  from the

time-translation invariance hypothesis); if  $\mathcal{O} = \omega_{k_1}(n_1) \omega_{k_2}(n_2)$ , then  $\mu[\mathcal{O}]$  is the probability that neurons  $k_1$  and  $k_2$  fire over the time span  $n_2 - n_1$ . Additionally,  $\mu[\omega_{k_1}(0) \omega_{k_2}(0)] - \mu[\omega_{k_1}(0)] \mu[\omega_{k_2}(0)]$  represents the instantaneous pairwise correlation between the neurons  $k_1$  and  $k_2$ .

There are several methods which allow the computation or estimation of  $\mu$ . In the following we shall assume that neural activity is described by a Markov process with memory depth  $D$  and positive time-translation invariant transition probabilities  $P[\omega(D) | \omega_0^{D-1}] > 0$ . From the assumption  $P[\omega(D) | \omega_0^{D-1}] > 0$ , this chain has a unique invariant probability  $\mu$  such that, for any  $n > D$ , and any block  $\omega_0^n$ :

$$\mu[\omega_0^n] = \prod_{l=0}^{n-D} P[\omega(D+l) | \omega_l^{D+l-1}] \mu[\omega_0^{D-1}]. \quad (2)$$

Therefore, knowing the transition probabilities (corresponding to blocks  $\omega_0^D$  of range  $D+1$ ) and  $\mu$  (which can be determined as well from the transition probabilities as exposed in section 2.3.1), the probability of larger blocks can be computed. Equation (2) makes explicit the role of memory in statistics of spike blocks, via the product of transition probabilities and the probability of the initial block  $\mu[\omega_0^{D-1}]$ .

On the opposite, if  $D = 0$ , the probability to have the spike pattern  $\omega(D)$  does not depend on the past activity of the network (memory-less case). In this case  $P[\omega(D+l) | \omega_l^{D+l-1}]$  becomes  $\mu[\omega(l)]$  and the probability (2) of a block becomes:

$$\mu[\omega_0^n] = \prod_{l=0}^n \mu[\omega(l)]. \quad (3)$$

Therefore, in the memory-less case, spikes occurring at different times are independent.

This emphasizes the deep difference between the case  $D = 0$  and the case  $D > 0$ .

#### 2.1.4 Empirical average

Let us assume that we are given an experimental raster of length  $T$ , such that  $\omega_0^{T-1}$ . The estimation of spikes statistics has to be done on this sample. In the context of the maximum entropy principle, where statistics is assumed to be time translation invariant, statistics of events is obtained via *time-average*. The time-average or empirical average of an observable  $\mathcal{O}$  in a raster  $\omega$  of length  $T$  is denoted by  $\pi_\omega^{(T)}[\mathcal{O}]$ . For example, if  $\mathcal{O} = \omega_k(0)$  the time-average  $\pi_\omega^{(T)}[\mathcal{O}] = \frac{1}{T-1} \sum_{n=0}^{T-1} \omega_k(n)$  is the firing rate of neuron  $k$ , estimated on the experimental raster  $\omega$ .

The empirical average is a random variable, depending on the raster  $\omega$ , as well as on the time length of the sample and it has Gaussian fluctuations whose amplitude tends to 0 as  $T \rightarrow +\infty$  like  $\frac{1}{\sqrt{T}}$ . This is the case, e.g. for the empirical averages obtained from several spike train acquired with several repetition.

### 2.1.5 Complexity of the set of spike blocks.

If one has  $N$  neurons and wants to consider spike block events within  $R$  time steps, one has  $2^{NR}$  possible states. For a reasonable Multi Electrodes Array (MEA) sample,  $N = 100$ ,  $R = 3$  (for a time lag of 30ms with a 10ms binning), this gives  $2^{300} \sim 4 \times 10^{180}$ , which is quite a bit more than the expected number of particles in the (visible) universe. Taking into account the huge number of states in the set of blocks, it is clear that any method requiring the extensive description of the state space will fail as  $NR$  grows. Additionally, while the accessible state space is huge, the *observed* state space (e.g. in an experimental raster) is rather small. For example, in a MEA raster for a retina experiment, the sample size  $T$  is about  $10^6 - 10^7$ , which is quite a bit less than  $2^{NR}$ . As a matter of fact, any reasonable estimation method must take this small-sample constraint into account. As we show in the next section, the maximum entropy principle and the related notion of Gibbs distributions allows us to take these aspects into account.

## 2.2 The maximum entropy principle

### 2.2.1 Motivations

Following 2.1 the goal is to find a probability distribution  $\mu$  such that:

- $\mu$  is inferred from an empirical raster  $\omega$ , by computing the empirical average of a set of ad hoc observables  $\mathcal{O}_k$ ,  $k = 1, \dots, K$ . One asks that the average of  $\mathcal{O}_k$  with respect to  $\mu$  satisfies:

$$\mu[\mathcal{O}_k] = \pi_\omega^{(T)}[\mathcal{O}_k], \quad k = 1, \dots, K. \quad (4)$$

The mean of  $\mathcal{O}_k$ , predicted by  $\mu$  is equal to the mean computed on the experimental raster.  $\mu$  is called a “model” in the sequel. The set of observables  $\mathcal{O}_k$  defines the model.

- $\mu$  has to be “as simple as possible”, with the least structure and a minimum number of tunable parameters. In the maximum entropy paradigm [28] these issues are (partly) solved by seeking a probability distribution  $\mu$  which maximizes the entropy under the constraints (4). The entropy is defined explicitly below (see eq. (5),(21)).

- From the knowledge of  $\mu$  one can compute the probability of arbitrary blocks (e.g. via eq. (2)) and the average of other observables than the  $\mathcal{O}_k$ s.

**Remark.**

Assume that we want to select observables  $\mathcal{O}_k$  in the set of all possible observables with range  $R$ . For  $N$  neurons there are  $2^{NR}$  possible choices. When  $NR$  increases, the number of possible observables will quickly exceed the number of samples available in the recording. Including all of them in the model would overfit the data. Therefore, one has to guide the choice of observables by additional criteria. We now review some of the criteria, which have been used by other authors.

### 2.2.2 Spatial models

In a seminal paper, Schneidman et al [56] aimed at unravelling the role of instantaneous pairwise correlations in retina spike trains. Although these correlations are weak, researchers investigated whether they play a more significant role in spike train statistics than firing rates.

Firing rates correspond to the average of observables of the form  $\omega_i(0)$ ,  $i = 1, \dots, N$  (the time index 0 comes from the assumed time-translation invariance) while instantaneous pairwise correlations correspond to averages of observables of the form  $\omega_i(0)\omega_j(0)$ ,  $1 \leq i < j \leq N$ . Analysing the role of pairwise correlations in spike train statistics, compared to firing rate, amounts therefore to comparing two models, defined by two different types of observables.

Note that all of these observables correspond to spatial events occurring at the same time. They give no information on how the spike patterns at a given time depend on the past activity. This situation corresponds to a memory-less model ( $D = 0$  in section 2.1.3), where transition probabilities do not depend on the past. As a consequence the sought probability  $\mu$  weights blocks of range 1, and the probability of blocks with larger range is given by (3): spike patterns at successive time steps are independent in spatial models.

In this case, the *entropy* of  $\mu$  is given by:

$$S(\mu) = - \sum_{\omega(0)} \mu[\omega(0)] \log \mu[\omega(0)]. \quad (5)$$

The natural log could be replaced by the logarithm in base 2.

Now, the maximum entropy principle of Jaynes [28] corresponds to seeking a probability  $\mu$  which maximizes  $S(\mu)$  under the constraints (4). It can be shown (see section 2.3 for the general statement) that this maximization problem is equivalent to the following Lagrange problem: maximising the quantity  $S(\mu) + \mu[\mathcal{H}_\beta]$  where  $\mathcal{H}_\beta$ , called a *potential* is given by:

$$\mathcal{H}_\beta = \sum_{k=1}^K \beta_k \mathcal{O}_k. \quad (6)$$

The  $\beta_k$ s are real numbers and free parameters.  $\mu[\mathcal{H}_\beta]$  is the average of  $\mathcal{H}_\beta$  with respect to  $\mu$ . Since  $\mathcal{H}_\beta$  is a linear combination of observables we have  $\mu[\mathcal{H}_\beta] = \sum_{k=1}^K \beta_k \mu[\mathcal{O}_k]$ . If the  $\mathcal{O}_k$ s have finite range and are  $> -\infty$  and if the  $\beta_k$ s are finite then it can be shown (see section 2.3) that there is only one probability  $\mu$ , depending on the  $\beta_k$ s, which solves the maximization problem. It is called a *Gibbs distribution*.

In this context ( $D = 0$ ) it reads:

$$\mu[\omega(0)] = \frac{e^{\mathcal{H}_\beta(\omega(0))}}{Z_\beta}, \quad (7)$$

where the normalization factor

$$Z_\beta = \sum_{\omega(0)} e^{\mathcal{H}_\beta(\omega(0))}, \quad (8)$$

is the so-called *partition function*.

To match (4) the parameters  $\beta_k$  have to be tuned. This can be done thanks to the following property of  $Z_\beta$ :

$$\mu[\mathcal{O}_k] = \frac{\partial \log Z_\beta}{\partial \beta_k}. \quad (9)$$

Thus the  $\beta_k$ s have to be tuned so that  $\mu$  matches (4) as well as (9). It turns out that  $\log Z_\beta$  is convex with respect to  $\beta_k$ s, so the problem has a unique solution.

Note that  $\log Z_\beta$  does not only allow us to obtain the averages of the observables, it also allows to characterize fluctuations. If a raster is distributed according to the Gibbs distribution (7), then, as pointed out in section 2.1.4, the empirical average of an observable has fluctuations. One can show that these fluctuations are Gaussian (Central limit theorem). The joint probability of  $\pi_\omega^{(T)}[\mathcal{O}_k]$ ,  $k = 1, \dots, K$  is Gaussian, with mean  $\mu[\mathcal{O}_k]$  given by (9) and covariance matrix  $\frac{\Sigma}{T}$  where the matrix  $\Sigma$  has entries:

$$\Sigma_{kl} = \frac{\partial^2 \log Z_\beta}{\partial \beta_k \partial \beta_l}. \quad (10)$$



Let us now discuss what this principle gives in the two cases considered by Schneidman et al.

(i) **Only firing rates are constrained.** Then:

$$\mathcal{H}_\beta(\omega(0)) = \sum_{k=1}^N \beta_k \omega_k(0).$$

It can be shown that the corresponding probability  $\mu$  is:

$$\mu[\omega(0)] = \prod_{k=1}^N \frac{e^{\beta_k \omega_k(0)}}{1 + e^{\beta_k}}.$$

Thus, the corresponding statistical model is such that spikes emitted by distinct neurons at the same time are independent. The parameter  $\beta_k$  is directly related to the firing rate  $r_k$  since  $r_k = \mu[\omega_k(0) = 1] = \frac{e^{\beta_k}}{1 + e^{\beta_k}}$ , so that we have:

$$\mu[\omega_0^n] = \prod_{l=0}^n \prod_{k=1}^N r_k^{\omega_k(l)} (1 - r_k)^{1 - \omega_k(l)},$$

the classical probability of coin tossing with independent probabilities (*Bernoulli model*). Thus, fixing only the rates as constraints, the maximum entropy principle leads to analyze spike statistics as if each spike were thrown randomly and independently, as with coin tossing. This is the “most random model”, which has the advantage of making as few hypothesis as possible. However, when only constrained with mean firing rates, the prediction of even small spike blocks in the retina was not successful. This was expected since this model assumes independence between neurons, an assumption that has been proven wrong in earlier studies (e.g. [51]).

(ii) **Firing rates and pairwise correlations are constrained.** In the second model, Schneidman et al constrained the maximum entropy model with both mean firing rates and instantaneous pairwise correlations between neurons. In this case,

$$\mathcal{H}_\beta(\omega(0)) = \sum_{k=1}^N \beta_k \omega_k(0) + \sum_{k,l=1}^N \beta_{kl} \omega_k(0) \omega_l(0).$$

Here the potential can be identified with the Hamiltonian of a magnetic system with binary spins. It is thus often called “Ising model”

in the spike train analysis literature, although the original Ising model has constant and positive couplings [22]. The corresponding statistical model is the least structured model respecting these first and second order pairwise instantaneous constraints. The number of parameters is of the order<sup>1</sup> of  $N^2$ , to be compared with the  $2^N$  possible patterns.

Schneidman et al showed that the Ising model model successfully predicts spatial patterns, a result which was confirmed by [62] (see [41] for a review). Other works have used the same method and found also a good prediction in cortical structure in vitro [70], in the visual cortex in vivo [80]. Later on, several authors considered higher order terms still corresponding to  $D = 0$  ([42, 56, 74, 19]). Note that these results have been obtained on relatively small subsets of neurons (usually groups of 10). An interesting challenge is to test how these results hold for larger subsets of neurons, and if other constraints have to be added [20](Tkacik et al, in preparation).

### 2.2.3 One time step spatio-temporal models and detailed balance

These models are only designed to predict the occurrence of “spatial” patterns, lying within one time bin. The use of spatial observables naturally leads to a time independence assumption where the probability of occurrence of a spatio-temporal pattern is given by the product (3). Tang et al. [70] tried to predict the temporal statistics of the neural activity with such a model and showed that it does not give a faithful description of temporal statistics. The idea to consider spatio-temporal observables then naturally emerges with the problem of generalising the probability eq. (7) to that case.

From the statistical mechanics point of view, a natural extension consists of considering the space of rasters  $\Omega$  as a lattice where one dimension is "space" (neurons index) and the other is time. The idea is then to consider a potential still of the form (6) but where the observables correspond to spatio-temporal events. We assume that  $\mathcal{H}_\beta$  has range  $R = D + 1$ ,  $0 \leq D < +\infty$ . The potential of a spike block  $\omega_0^n$ ,  $n \geq D$  is:

$$\mathcal{H}_\beta(\omega_0^n) = \sum_{l=0}^{n-D} \mathcal{H}_\beta(\omega_l^{D+l}) \quad (11)$$

On this basis, restricting to the case where  $D = 1$  (one time step memory depth) Marre et al have proposed in [37] to construct a Markov chain, where transition probabilities  $P[\omega(l+1) | \omega(l)]$  are proportional to  $e^{\mathcal{H}_\beta(\omega_l^{l+1})}$ . If

---

<sup>1</sup>Most approaches assumes moreover that the pairwise coefficients are symmetric  $\beta_{kl} = \beta_{lk}$  which divides the number of parameters by 2.

$\mu$  is the invariant probability of that chain, the application of (2) leads to probability of blocks  $\mu[\omega_0^n]$ , proportional to  $e^{\mathcal{H}_\beta(\omega_0^n)}$ : the probability of a block is proportional to the exponential of its potential ("energy"). This approach is therefore quite natural from the statistical mechanics point of view.

The main problem, however, is "what is the proportionality coefficient?" As shown in [37], the normalization of conditional probabilities does not reduce to the mere division by a constant partition function. This normalization factor is itself dependent on the past activity.

To overcome this dependency, Marre et al assumed that the activity respected a detailed balance. In this particular case, it can be shown that the normalization factor becomes, again, a constant. But this is an important reduction that could have implications for the interpretation of the data: for example, with this simplification, it is not possible to give an account of assymmetric cross-correlograms.

## 2.3 General spatio-temporal models

We now present the general formalism which allows to solve the variational problem "maximising entropy under spatio-temporal constraints". This approach is rigorous and the normalization problem is resolved without requiring additional assumptions such as detailed balance. At the end of this section, we briefly discuss other approaches considering spatio-temporal statistics and their relations to potentials of the form (6).

### 2.3.1 Constructing the Markov Chain

In this section we show how one can generate a Markov chain where transition probabilities are proportional to  $e^{\mathcal{H}_\beta(\omega_l^{l+D})}$ , for a potential  $\mathcal{H}_\beta$  corresponding to spatio-temporal events. We also solve the normalization problem. This construction is well known and is based on the so-called transfer matrix (see e.g. [22] for a presentation in the context of statistical physics; [46] for a presentation in the context of ergodic theory and [78] for a presentation in the context of spike trains analysis).

This matrix is constructed as follows. Consider two spike blocks  $w_1, w_2$  of range  $D \geq 1$ . The transition  $w_1 \rightarrow w_2$  is *legal* if  $w_1$  has the form  $\omega(0)\omega_1^{D-1}$  and  $w_2$  has the form  $\omega_1^{D-1}\omega(D)$ . The vectors  $\omega(0), \omega(D)$  are arbitrary but the block  $\omega_1^{D-1}$  is common. Here is an example of a legal transition :

$$w_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}; w_2 = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}.$$

Here is an example of a forbidden transition

$$w_1 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}; w_2 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Any block  $\omega_0^D$  of range  $R = D + 1$  can be viewed as a legal transition from the block  $w_1 = \omega_0^{D-1}$  to the block  $w_2 = \omega_1^D$  and in this case we write  $\omega_0^D \sim w_1 w_2$ .

The *transfer matrix*  $\mathcal{L}$  is defined as:

$$\mathcal{L}_{w_1, w_2} = \begin{cases} e^{\mathcal{H}_\beta(\omega_0^D)} & \text{if } w_1, w_2 \text{ is legal with } \omega_0^D \sim w_1 w_2 \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

From the matrix  $\mathcal{L}$  the transition matrix of a Markov chain can be constructed, as we now show. Since observables are assumed to be bounded from below,  $\mathcal{H}_\beta(\omega_0^D) > -\infty$ , thus  $e^{\mathcal{H}_\beta(\omega_0^D)} > 0$  for each legal transition. As a consequence of the Perron-Frobenius theorem [21, 57],  $\mathcal{L}$  has a unique real positive eigenvalue  $s_\beta$ , strictly larger than the modulus of the other eigenvalues (with a positive gap), and with right,  $R$ , and left,  $L$ , eigenvectors:  $\mathcal{L}R = s_\beta R$ ,  $L\mathcal{L} = s_\beta L$ , or, equivalently<sup>2</sup>:

$$\sum_{\omega(D) \in \{0,1\}^N} e^{\mathcal{H}_\beta(\omega_0^D)} R(\omega_1^D) = s_\beta R(\omega_0^{D-1});$$

$$\sum_{\omega(0) \in \{0,1\}^N} L(\omega_0^{D-1}) e^{\mathcal{H}_\beta(\omega_0^D)} = s_\beta L(\omega_1^D).$$

These eigenvectors have strictly positive entries  $R(\cdot) > 0$ ,  $L(\cdot) > 0$ , functions of blocks of range  $D$ . They can be chosen so that the scalar product  $\langle L, R \rangle = 1$ . We define:

$$\mathcal{P}(\mathcal{H}_\beta) = \log s_\beta. \quad (13)$$

called "topological pressure". We discuss the origin of this term and its properties in the section 2.3.2.

---

<sup>2</sup>The right eigenvector  $R$  has  $2^{ND}$  entries  $R_w$  corresponding to blocks of range  $D$ . It obeys  $\sum_{w_2} \mathcal{L}_{w_1 w_2} R_{w_2} = s_\beta R_{w_1}$ , where  $w_1 = \omega_0^{D-1}$  and where the sum runs over blocks  $w_2 = \omega_1^D$ . Since  $\mathcal{L}_{w_1 w_2}$  is non zero only if the entries  $w_1, w_2$  have the block  $\omega_1^{D-1}$  in common, and since the right hand side ( $s_\beta R_{w_1}$ ) fixes the value of  $w_1$ , this sum holds in fact on all possible values of  $\omega(D)$ . The notation  $R_w$ , although natural, does not make explicit the block involved. This is problematic when one wants to handle equations such as (19). As a consequence, we prefer to use the notation  $R(\text{block})$  to make explicit this dependence. The same remark holds mutatis mutandis for the left eigenvector.

To define a Markov chain from the transfer matrix  $\mathcal{L}$  (eq. 12 ) we introduce the *normalised potential*:

$$\phi(\omega_0^D) = \mathcal{H}_\beta(\omega_0^D) - \mathcal{G}_\beta(\omega_0^D) \quad (14)$$

with:

$$\mathcal{G}_\beta(\omega_0^D) = \log R(\omega_0^{D-1}) - \log R(\omega_1^D) + \log s_\beta, \quad (15)$$

and a family of transition probabilities:

$$P[\omega(D) \mid \omega_0^{D-1}] \stackrel{\text{def}}{=} e^{\phi(\omega_0^D)} > 0. \quad (16)$$

These transition probabilities define a Markov chain which admits a unique invariant probability:

$$\mu(\omega_0^{D-1}) = R(\omega_0^{D-1}) L(\omega_0^{D-1}). \quad (17)$$

From the general form of block probabilities (2) the probability of blocks of depth  $n \geq D$  is, in this case :

$$\mu[\omega_0^n] = e^{\sum_{t=0}^{n-D} \phi(\omega_t^{D+t})} \mu[\omega_0^{D-1}]. \quad (18)$$

thus, from (17),(14),(15):

$$\mu[\omega_0^n] = \frac{e^{\mathcal{H}_\beta(\omega_0^n)}}{s_\beta^{n-D+1}} R(\omega_{n-D+1}^n) L(\omega_0^{D-1}), \quad (19)$$

where  $\mathcal{H}_\beta(\omega_0^n)$  is given by (11).

### 2.3.2 Remarks

1. We have been able to compute the probability of any blocks  $\omega_0^n$ . It is proportional to  $e^{\mathcal{H}_\beta(\omega_0^n)}$  and the proportionality factor has been computed. In the general case of spatio-temporal events, it depends on  $\omega_0^{D-1}$  and  $\omega_{n-D+1}^n$ .

The same arises in statistical mechanics when dealing with boundary conditions. The forms (18), (19), remind Gibbs distributions on spin lattices, with lattice translation-invariant probability distributions given specific boundary conditions. Given a spin-potential of spatial-range  $n$ , the probability of a spin block depends upon the state of the spin block, as well as spins states in a neighbourhood of that block. The conditional probability of this block given a fixed neighbourhood is the exponential of the energy characterizing physical interactions, within

the block, as well as interactions with the boundaries. In (18), spins are replaced by spiking patterns; space is replaced by time. Spatial boundary conditions are here replaced by the dependence upon  $\omega_0^{D-1}$  and  $\omega_{n-D+1}^n$ .

As a consequence, as soon as one is dealing with spatio-temporal events the normalization of conditional probabilities does not reduce to the mere division by:

$$Z_n = \sum_{\omega_0^n} e^{\mathcal{H}_\beta(\omega_0^n)}, \quad (20)$$

as easily checked in (19).

2. The topological pressure obeys nevertheless:

$$\mathcal{P}(\mathcal{H}_\beta) = \lim_{n \rightarrow +\infty} \frac{1}{n} \log Z_n,$$

and is analogue to a thermodynamic potential density (free energy, free enthalpy, pressure). This analogy is also clear in the variational principle (23) below. To our best knowledge the term "topological pressure" has its roots in the thermodynamic formalism of hyperbolic (chaotic) maps [55, 46, 5]. In this context, this function can be computed as the grand potential of the grand canonical ensemble, as a cycle expansion over unstable periodic orbits. It is therefore equivalent to a pressure<sup>3</sup> depending on topological properties (periodic orbits).

3. In the case  $D = 0$  the Gibbs distribution reduces to (7). One can indeed easily show that:

$$\exp \mathcal{G}_\beta = s_\beta = \sum_{\omega(0)} e^{\mathcal{H}_\beta(\omega(0))} = Z_\beta,$$

where  $Z_\beta$  is the partition function (8). Additionally, since spike patterns occurring at distinct time are independent in the  $D = 0$  case,  $Z_n$  in (20) can be written as  $Z_n = Z_\beta^n$  so that  $\mathcal{P}(\mathcal{H}_\beta) = \log Z_\beta$ .

4. In the general case of spatio-temporal constraints, the normalization requires the consideration of normalizing function  $\mathcal{G}_\beta$  *depending as well on the blocks  $\omega_0^D$* . Thus, in addition to function  $\mathcal{H}_\beta$  normalization introduces a second *function* of spike blocks. This increases consequently the complexity of Gibbs potentials and Gibbs distributions compared to the spatial ( $D = 0$ ) case where  $\mathcal{G}_\beta$  reduces to a constant.

---

<sup>3</sup>The grand potential  $\Phi$  obeys  $\Phi = -PV$ , where  $P$  is the physical pressure and  $V$  the volume. Therefore, the grand potential density is (minus) the pressure.

### 2.3.3 The maximum entropy principle

We now show that the probability distribution defined this way solves the variational problem “maximising entropy under constraints”.

We define the *entropy rate* (or Kolmogorov-Sinai entropy):

$$h[\mu] = - \limsup_{n \rightarrow \infty} \frac{1}{n+1} \sum_{\omega_0^n} \mu[\omega_0^n] \log \mu[\omega_0^n], \quad (21)$$

where the sum holds over all possible blocks  $\omega_0^n$ . Note, that in the case of a Markov chain  $h[\mu]$  also reads [17]:

$$h[\mu] = - \sum_{\omega_0^D} \mu[\omega_0^D] P[\omega(D) | \omega_0^{D-1}] \log P[\omega(D) | \omega_0^{D-1}], \quad (22)$$

whereas, when  $D = 0$ ,  $h[\mu]$  reduces to the definition (5).

As a general result from ergodic theory [55, 30, 13] and mathematical statistical physics [22], there is a unique<sup>4</sup> probability distribution  $\mu$  such that [55, 30, 13]:

$$\mathcal{P}[\mathcal{H}_\beta] = \sup_{\nu \in \mathcal{M}_{inv}} (h[\nu] + \nu[\mathcal{H}_\beta]) = h[\mu] + \mu[\mathcal{H}_\beta], \quad (23)$$

where  $\mathcal{P}[\mathcal{H}_\beta]$  is given by (13).  $\mathcal{M}_{inv}$  is the set of all possible time-translation invariant probabilities on the set of rasters with  $N$  neurons and  $\nu[\mathcal{H}_\beta] = \sum_{\omega_0^D} \mathcal{H}_\beta(\omega_0^D) \nu(\omega_0^D)$  is the average value of  $\mathcal{H}_\beta$  with respect to the probability  $\nu$ .

Looking at the second equality, the variational principle (23) selects, among all possible probabilities  $\nu$ , a unique one realizing the supremum. This is exactly the invariant distribution of the Markov chain and is the sought Gibbs distribution. It is clear from (23) that the topological pressure is the formal analogue to a thermodynamic potential density, where  $\mathcal{H}_\beta$  somewhat fixes the "ensemble":  $\nu[\mathcal{H}_\beta] = \sum_{k=1}^K \beta_k \nu[\mathcal{O}_k]$  plays the role of  $\beta E$  (canonical ensemble),  $\beta E - \mu N$  (grand canonical ensemble), ... in thermodynamics [5].

### 2.3.4 Inferring the parameters $\beta_k$

The inverse problem of finding the values of  $\beta_k$ s from the observables average measured on the data is a hard problem with no exact analytical solution.

---

<sup>4</sup>The result is straightforward here since we consider bounded potentials with finite range.

However, in the context of spatial models with pairwise interactions the wisdom, coming from statistical physics and especially Ising model and spin-glasses, as well as from the Boltzmann machine learning community, can be used. As a consequence, in this context, several strategies were proposed. Ackley et al [2] proposed a technique to estimate the parameters of a Boltzmann machine. This technique is effective for small networks but it is time consuming. In practice, the time necessary to learn the parameters increases exponentially with the number of units. To speed up the parameters estimation, analytical approximations of the inverse problem have been proposed, which express the parameters  $\beta_k$  as a nonlinear function of the correlations of the activity (see for example [69], [52], [58], [3], [54], [2], [26], [29]).

These methods do not give an exact result, but are computationally fast. We do not pretend to review all of them here, but we quote a few prominent examples. In [58], Sessak and Monasson proposed a systematic small-correlation expansion to solve the inverse Ising problem. They were able to compute couplings up to the third order in the correlations for generic magnetizations, and to the seventh order in the case of zero magnetizations. Their resulting expansion outperforms existing algorithms on the Sherrington-Kirkpatrick spin-glass model [61].

Based on a high-field expansion of the Ising thermodynamic potential, Cocco et al [15] designed an algorithm to calculate the parameters in a time polynomial with  $N$ , where the couplings are expressed as a weighted sum over the power of the correlations. They did not obtain a closed analytical expression, but their algorithm could run in a time that was polynomial in the number of neurons.

Other methods, based on Thouless-Anderson-Palmer equations [72] and linear response [29], or information geometry [69], initially proposed in the field of spin-glasses, have been adapted and applied to spike train analysis (see e.g. the work done by Roudi and collaborators [53]).

The success of these approximations depends on the dataset, and there is no a priori guarantee about their efficiency at finding the right values of the parameters. However, by getting closer to the correct solution, they can potentially speed up the convergence of the learning by starting with a seed much closer to the real solution than if taking a random starting point.

Note also that all the techniques mentioned above have been designed for the case where there is no temporal interaction (except [15, 53] which are discussed in the section 2.3.5). Now, we explain how the parameters estimation can be done in the spatio-temporal models.

In the general case parameters  $\beta_k$ s can be determined thanks to the following properties.



- $\mathcal{P}[\mathcal{H}_\beta]$  is a log generating function of cumulants. First:

$$\frac{\partial \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k} = \mu[\mathcal{O}_k]. \quad (24)$$

This is an extension of (9) to the time-dependent case.

- Second:

$$\frac{\partial^2 \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k \partial \beta_l} = \frac{\partial \mu[\mathcal{O}_k]}{\partial \beta_l} = \sum_{n=0}^{+\infty} C_{\mathcal{O}_k \mathcal{O}_l}(n), \quad (25)$$

where  $C_{\mathcal{O}_k \mathcal{O}_l}(n)$  is the correlation function between the two observables  $\mathcal{O}_k$  and  $\mathcal{O}_l$  at time  $n$ . Note that correlation functions decay exponentially fast whenever  $\mathcal{H}_\beta$  has finite range. So that  $\sum_{n=0}^{+\infty} C_{\mathcal{O}_k \mathcal{O}_l}(n) < +\infty$ .

Eq. (25) characterizes the variation in the average value of  $\mathcal{O}_k$  when varying  $\beta_l$  (linear response). The corresponding matrix is a susceptibility matrix. It controls the Gaussian fluctuations of observables around their mean (central limit theorem) [55, 46, 13]. This is the generalization of (10) to the time dependent case. As a particular case, the fluctuations of the empirical average  $\pi_\omega^{(T)}[\mathcal{O}_k]$  of  $\mathcal{O}_k$  around its mean  $\mu[\mathcal{O}_k]$  are Gaussian with a mean-square deviation  $\frac{\sqrt{\mu[\mathcal{O}_k](1-\mu[\mathcal{O}_k])}}{\sqrt{T}}$ .

It is clear that the structure of the linear response in the case of spatio-temporal constraints is quite a bit more complex than the case  $D = 0$  (see eq. (10)). Actually, for  $D = 0$ , all correlations  $C_{\mathcal{O}_k \mathcal{O}_l}(n)$  vanish for  $n > 0$  (distinct times are independent).

- $\mathcal{P}(\mathcal{H}_\beta)$  is a convex function of  $\beta$ . As a consequence, if there is a set of  $\beta$  value,  $\beta^*$ , such that .

$$\frac{\partial \mathcal{P}[\mathcal{H}_\beta]}{\partial \beta_k^*} = \mu[\mathcal{O}_k] = C_k, \quad (26)$$

then this set is unique. Thus, the solution of the variational problem (23) is unique.

Basically, eq. (24), (25), 26, tell us that techniques based on free energy expansion in spatial models can be extended as well to spatio-temporal cases, where the free energy is replaced by the topological pressure. Obviously, estimating (not to speak of computing) the topological pressure can be a formidable task. Although, the transfer matrix technique allows the computation of the topological pressure, the use of this method for large  $N$

is hopeless (see section 3.1). However, techniques based on periodic orbit expansion and zeta functions, could be useful [46]. Additionally, cumulant expansions of the pressure, eq. (24) and (25) corresponding to the two first orders, suggest that extension of methods based on free energy expansion could be used. In addition to the works quoted above, we also think of constraint satisfaction problems by Mézard and Mora [39] and approaches based on Bethe free energy [79]. Finally, as we checked, the properties of spatio-temporal Gibbs distributions allows to extend the parameters estimation methods developed for the spatial case in [18, 8] to spatio-temporal distributions (to be published).

### 2.3.5 Other spatio-temporal models

Here we shortly review alternative spatio-temporal models. We essentially refer to approaches attempting to construct a Markov chain and related invariant probability by proposing a specific form for the transition probabilities.

A prominent example is provided by the so-called Linear-Nonlinear (LN) models and Generalized Linear Models (GLM) [6, 38, 45, 75, 49, 47, 4, 48]. Shortly, the idea is to model spike statistics by a point process where the instantaneous firing rate of a neuron is a nonlinear function of the past network activity including feedbacks and interaction between neurons [64]. This model has been applied in a wide variety of experimental settings [7, 14, 71, 9, 44, 75, 47]. Typically, referring e.g. to [4], the rate  $r_i$  has the form (adapting to our notations):

$$r_i = f \left[ b_i + K_i \cdot x + \sum_j H_{ij} \right] \quad (27)$$

where the kernel  $K_i$  represents the  $i$ -th cell's linear receptive field and  $x$  an input.  $H_{ij}$  characterizes the effect of spikes emitted in the past by pre-synaptic neuron  $j$  on post-synaptic neuron  $i$ . In this approach, neurons are assumed to be conditionally-independent given the past. The probability to have a given spike-response to a stimulus, given the past activity of the network, reads as the product of firing rates (see e.g. eq. 2.4 in [4]).

In [4] the authors use several Monte Carlo approaches to learn the parameters of the model for a Bayesian decoding of the rasters. Comparing to the method presented in the previous sections, the main advantages of the GLM is: (i) The transition probability is known (postulated) from the beginning and does not require the heavy normalization imposed by potentials of the form (6); (ii) The model parameters have a neurophysiological interpretation,

and their number grows at most as a power law in the number of neurons, as opposite to (6), where the parameters are delicate to interpret and whose number can become quite large, depending on the set of constraints.

Note however that a model of the form (27) can be written as well in the form (6): this is a straightforward consequence of the Hammersley-Clifford theorem [23]. The parameters  $\beta_k$  in (6) are then nonlinear functions of the parameters in (27) (see [11] for an example).

The main drawback of this approach is the assumption of conditional independence between neurons: neurons are assumed independent at time  $t$  when the past, which appears in the function  $H_{ij}$  in (27), is given, and the probability of a spiking pattern at time  $t$  is the product of neurons firing rates. On the opposite, the maximal entropy principle does not require this assumption.

It is interesting to remark that the conditional independence assumption can be rigorously justified in conductance based Integrate and Fire models [11, 12] and the form of the function  $f$  can be explicitly found (this a sigmoid function instead of an exponential as usually postulated in GLM). This result holds true if only chemical synapses are involved (this is also implicit in the kernel form  $H_{ij}$  in (27) [4]), but conditional independence breaks down for example as soon as electric synapses (gap junctions) are involved: this can be mathematically shown in conductance based Integrate and Fire models [16]. Note that in this case, a large part of correlations is due dynamical interactions between neurons: as a consequence they persist even if there is no shared input.

Recently, Macke et al. [36] extended the GLM model to fix the lack of instantaneous correlations between neurons in the GLM. They added a common input function that has a linear temporal dynamics. However, one of the disadvantages of this technique is that its likelihood is not unimodal, and thus that computationally expensive Expectation-Maximization algorithms have to be used to fit parameters.

The GLM model is usually used to model both the stimulus-response dependence as well as the interaction between neurons, while the MaxEnt models usually focus on the latter (but see [73]).

To finish this subsection, we would like to quote two important works dealing with spatio-temporal events too. First, In [15] Cocco and co-workers consider retinal ganglion cells spiking activity with a dual approach: on one hand they consider an Ising model (and higher order spatial terms) where they propose an inverse method based on a cluster expansion to find efficiently the coupling in Ising model from data; on the other hand, they

consider the problem of finding the parameters (synaptic couplings) in a Integrate and Fire model with noise, from its spike trains. In the weak noise limit the conditional probability of a spiking pattern given the past is given by a least action principle. This probability is a Gibbs distribution whose normalized potential is characterized by the action computed over an optimal path. This second approach allows the characterization of spatio temporal events. Especially it gives a very good fit of the cross-correlograms.

Second, in [53], the authors consider a one step memory Markov chain where the conditional probability has a time-dependent potential of Ising type. Adapting a Thouless-Anderson-Palmer [72] approach used formerly in the Sherrington-Kirkpatrick mean-field model of spin glasses [61] they propose an inversion algorithm to find the model-parameters. As in the GLM their model assumes conditional independence given the past (see eq. (1) in [53]).

## 2.4 Comparing models

Solving equations (26) provides an optimal choice for the Gibbs distribution  $\mu$ , *given the observables*  $\mathcal{O}_k$ . However, changing the set of observables provides distinct Gibbs distributions, which does not approximate the hidden probability with the same accuracy. We need here a way to quantify the “distance” between the “model” (the Gibbs distribution fixed by the set of observables) and the exact, hidden, probability  $\mu^{(*)}$ . Here are several criteria of comparison.

### 2.4.1 Kullback-Leibler divergence

The Kullback-Leibler divergence between  $\mu, \mu^{(*)}$  is given by:

$$d(\mu^{(*)}, \mu) = \limsup_{n \rightarrow \infty} \frac{1}{n+1} \sum_{\omega_0^n} \mu^{(*)}[\omega_0^n] \log \left[ \frac{\mu^{(*)}[\omega_0^n]}{\mu[\omega_0^n]} \right], \quad (28)$$

which provides some notion of asymmetric “distance” between  $\mu$  and  $\mu^{(*)}$ . The KL divergence accounts for discrepancy between the predicted probability  $\mu[\omega_0^n]$  and the exact probability  $\mu^{(*)}[\omega_0^n]$  for all blocks of range  $n$ .

This quantity is not numerically computable from (28). However, for  $\mu$  a Gibbs distribution and  $\mu^{(*)}$  a time-translation invariant probability, the following holds:

$$d_{KL}(\mu^{(*)}, \mu) = \mathcal{P}[\mathcal{H}_\beta] - \mu^{(*)}[\mathcal{H}_\beta] - h(\mu^{(*)}).$$

The topological pressure  $\mathcal{P}[\mathcal{H}_\beta]$  is given by (13) while  $\mu^{(*)}[\mathcal{H}_\beta]$  is estimated by  $\pi_\omega^{(T)}[\mathcal{H}_\beta] = \sum_{k=1}^K \beta_k \pi_\omega^{(T)}[\mathcal{O}_k] = \sum_{k=1}^K \beta_k C_k$ .

Since  $\mu^{(*)}$  is unknown,  $h(\mu^{(*)})$  is unknown, and can only be estimated from data, i.e. one estimates the entropy of the empirical probability,  $h(\pi_\omega^{(T)})$ . There exist efficient methods for that. Note that the entropy of a Markov chain is readily given by eq. (22), so the entropy  $h(\pi_\omega^{(T)})$  is obtained by replacing the exact probability  $P$  in eq. (22), by the empirical probability  $h(\pi_\omega^{(T)})$ . As  $T \rightarrow +\infty$ ,  $h(\pi_\omega^{(T)}) \rightarrow h(\mu^{(*)})$ , at exponential rate<sup>5</sup>. For finite  $T$  finite size corrections exist, see e.g. Strong et al. [68]. In figure 1 is plotted an example. For a potential  $\mathcal{H}_\beta$  with  $N = 5$  neurons and range  $R = 2$ , containing all possible observables, we have plotted the difference between the exact probability (known from (22) and the explicit form (16), (17) of transition probabilities and invariant probability), and the approached entropy  $h(\pi_\omega^{(T)})$  obtained by replacing the exact probability  $P$  by the empirical probability  $h(\pi_\omega^{(T)})$ , as a function of raster size  $T$ . We have also plotted the finite corrections method proposed by Strong et al. in [68].

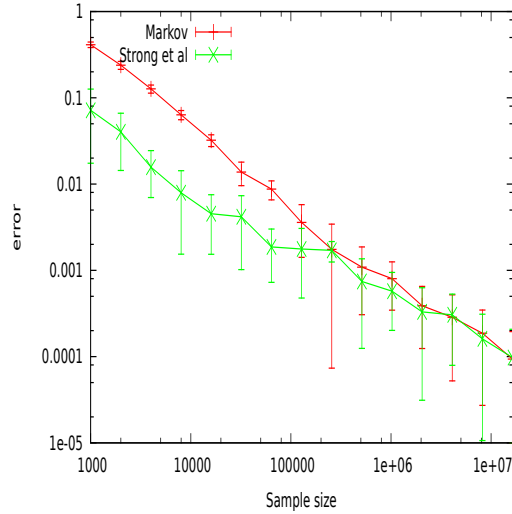


Figure 1: Difference between the exact probability and the approached entropy  $h(\pi_\omega^{(T)})$ , as a function of raster size  $T$ . The potential of test includes all the possible observables where weights are set as random values.

Now, if one wants to compare how two Gibbs distributions  $\mu_1, \mu_2$  ap-

<sup>5</sup>The rate is given by the spectral gap of the transfer matrix: the difference between the largest eigenvalue (it is real and positive), and the modulus of the second largest eigenvalue (in modulus).

proximate data, one compares the divergence  $d_{KL}(\mu^{(*)}, \mu_1)$ ,  $d_{KL}(\mu^{(*)}, \mu_2)$  where  $h(\mu^{(*)})$  is independent of the model choice. Therefore, the comparison of two models can be done without computing  $h(\mu^{(*)})$ .

#### 2.4.2 Comparison of observables average

Another criterion, easier to compute, is to compare the expected value of the observables average,  $\mu^{(*)}[\mathcal{O}_k]$ , known from (24) to the empirical average  $\pi_\omega^{(T)}[\mathcal{O}_k]$ . Error bars are expected to follow the central limit theorem where fluctuations are given by eq. (25). Examples are given in Fig. 2. Note that the comparison of observables average is less discriminant than minimizing the Kullback-Leibler divergence, since there are infinitely many possible models matching the observables average.

### 3 Monte-Carlo method for spatio-temporal Gibbs distribution

#### 3.1 The advantages and limits of transfer matrix method

The advantage of the transfer matrix technique method is that it is mathematically exact: given a potential  $\mathcal{H}_\beta$ , it gives the Gibbs distribution and topological pressure without computing a partition function; given the parametric form (6) where the parameters  $\beta_k$ s has to be determined (“learned”), it provides the unique solution. On numerical grounds, this method provides an optimal estimation, in the limits of the error made when observing the observables empirically, this error being characterized by the central limit theorem. Its main drawback is that the transfer matrix  $\mathcal{L}$  has  $2^{NR}$  entries ! Although, most of those entries are zero ( $2^N$  non zero entries per row, thanks to the compatibility conditions) it is simply too huge to handle cases where  $NR > 24$ .

Focusing thus on the huge number of states in the set of blocks, it is clear that any method requiring the extensive description of the phase space fails as  $NR$  grows. Additionally, while the accessible phase space is huge, the *observed* phase space (e.g. in an experimental raster) is rather small. Several strategies exist to avoid the extensive description of the phase space. Here, we propose an approach based on Monte-Carlo sampling.

The idea is the following. Given a potential  $\mathcal{H}_\beta$  we find a strategy to approximately compute the average  $\mu[\mathcal{O}_k]$  of observables  $\mathcal{O}_k$  under the Gibbs distribution  $\mu$ , using a statistical Monte-Carlo sampling of the phase space. For that purpose, the algorithm generates a raster following the statistics

defined by the potential  $\mathcal{H}_\beta$ , and computes the observables on this artificial raster. Thanks to the estimation of the observables, the parameters of the model ( $\beta_k$ s) can be found by modifying their values to minimize iteratively the distance between the values of the observables estimated on the real raster, and the values estimated with the Monte-Carlo sampling. Powerful algorithms exist for that, taking into account the uncertainty on the empirical averages ruled by the central limit theorem ([18] and [8]).

### 3.2 The Monte-Carlo-Hastings algorithm

The Monte-Carlo-Hastings method consists in sampling a target probability distribution  $\mu$  by constructing a Markov chain whose invariant probability is  $\mu$  [25]. The transition probability of this Markov chain, between two states  $\omega^{(1)}$  and  $\omega^{(2)}$  is:

$$P[\omega^{(1)}|\omega^{(2)}] = \min\left(\frac{Q(\omega^{(1)}|\omega^{(2)})}{Q(\omega^{(2)}|\omega^{(1)})} \frac{\mu[\omega^{(2)}]}{\mu[\omega^{(1)}]}, 1\right). \quad (29)$$

The function  $Q()$  can have different forms, allowing in particular to speed-up the convergence rate of the algorithm. Such specific forms are highly dependent on the form of  $\mathcal{H}_\beta$ , and there is no general recipe to determine  $Q$ , given  $\mathcal{H}_\beta$ . The contribution of  $Q$  cancels in (29) whenever  $Q$  is symmetric ( $Q(\omega|\omega') = Q(\omega'|\omega)$ ). We make this assumption in the sequel. Practically, we take  $Q$  as the uniform distribution corresponding to flipping one spike at each iteration of the method.

In classical Monte-Carlo approaches in statistical physics, the normalization factor of the Gibbs distribution, the partition function, cancels when computing the ratio of two blocks probabilities  $\frac{\mu[\omega^{(2)}]}{\mu[\omega^{(1)}]}$ . The situation is different in the presence of spatio temporal constraints, as shown in eq. (19): “boundary terms”  $L(\omega_0^{D-1})$ ,  $R(\omega_{n-D+1}^n)$  remain. Actually, the same would hold in statistical physics problem with spatial interactions if one were to compare the probability of bulk spin-chains with distinct boundary conditions.

This problem can however be circumvented thanks to the following remarks:

1. If one compares the probability of two blocks  $\omega^{(1)}, \omega^{(2)}$  of range  $n \geq 2D+1$ , with  $\omega_0^{D-1,(1)} = \omega_0^{D-1,(2)}$  and  $\omega_{n-D+1}^{n,(1)} = \omega_{n-D+1}^{n,(2)}$  then (19) reads:

$$\frac{\mu[\omega_0^{n,(2)}]}{\mu[\omega_0^{n,(1)}]} = e^{\Delta\mathcal{H}_\beta(\omega^{(1)}, \omega^{(2)}, 0, n)}$$

with

$$\Delta \mathcal{H}_\beta(\omega^{(1)}, \omega^{(2)}, 0, n) = \mathcal{H}_\beta \left( \omega_0^{n,(1)} \right) - \mathcal{H}_\beta \left( \omega_0^{n,(2)} \right).$$

Thus, the Monte-Carlo transition probability (29) is only expressed as a difference of potential of the two blocks.

2.  $\Delta \mathcal{H}_\beta(\omega^{(1)}, \omega^{(2)}, 0, n) = \sum_{k=1}^K \beta_k \Delta \mathcal{O}_k(\omega^{(1)}, \omega^{(2)}, 0, n)$ , with:

$$\Delta \mathcal{O}_k(\omega^{(1)}, \omega^{(2)}, 0, n) = \sum_{l=0}^{n-D} \left[ \mathcal{O}_k \left( \omega_l^{D+l,(2)} \right) - \mathcal{O}_k \left( \omega_l^{D+l,(1)} \right) \right].$$

Since the  $\mathcal{O}_k$ s are monomials, many terms  $\mathcal{O}_k(\omega_l'^{l+D}) - \mathcal{O}_k(\omega_l^{l+D})$  cancel. Assuming that we flip a spike at position  $(k, t)$ ,  $k \in \{1, \dots, N\}$ ,  $t \in \{D, n-D\}$ , we have indeed:

$$\Delta \mathcal{O}_k(\omega^{(1)}, \omega^{(2)}, 0, n) = \sum_{l=t-D}^t \left[ \mathcal{O}_k \left( \omega_l^{D+l,(2)} \right) - \mathcal{O}_k \left( \omega_l^{D+l,(1)} \right) \right]$$

Since the difference  $\mathcal{O}_k \left( \omega_l^{D+l,(2)} \right) - \mathcal{O}_k \left( \omega_l^{D+l,(1)} \right) \in \{-1, 0, 1\}$ , the computational cost of  $\Delta \mathcal{O}_k(\omega^{(1)}, \omega^{(2)}, 0, n)$  is minimal if one makes a list of monomials affected by the flip of spike  $(k, r)$ ,  $r = 0, \dots, D$ .

### 3.3 Convergence rate

The goal of Monte-Carlo-Hastings algorithm is to generate a sample of a target probability obtained by iteration of the Markov chain defined by eq. (29). In our case, this sample is a raster  $\omega_0^{T-1}$ , distributed according to a Gibbs distribution  $\mu$ . Call  $N_{flip}$  the number of iterations (“flips” in our case) of the Monte-Carlo algorithm. As  $N_{flip} \rightarrow +\infty$  the probability that the algorithm generates a raster  $\omega_0^{T-1}$  tends to  $\mu[\omega_0^{T-1}]$ . Equivalently, if one generates  $N_{seed}$  rasters and denote  $\#(\omega_0^{T-1})$  the number of occurrences of a specific bloc  $\omega_0^{T-1}$ , then:

$$\lim_{N_{seed} \rightarrow +\infty} \lim_{N_{flip} \rightarrow +\infty} \frac{\#(\omega_0^{T-1})}{N_{seed}} = \mu[\omega_0^{T-1}].$$

The convergence is typically exponential with a rate depending on  $\mathcal{H}_\beta$ .

Now, the goal here is to use a Monte-Carlo raster to estimate  $\mu[\mathcal{O}_k]$  by performing the empirical average  $\pi_\omega^{(T)}[\mathcal{O}_k]$  on that raster. However, as explained in section 2.1.4, even if the raster is distributed according to  $\mu$  (corresponding thus to taking the limit  $N_{flip} \rightarrow +\infty$ ) the empirical average



$\pi_\omega^{(T)}[\mathcal{O}_k]$  is not equal to  $\mu[\mathcal{O}_k]$ , it converges to  $\mu[\mathcal{O}_k]$  as  $T \rightarrow +\infty$ , with an exponential rate (see footnote 5). More precisely, the probability that the difference  $\left| \pi_\omega^{(T)}[\mathcal{O}_k] - \mu[\mathcal{O}_k] \right|$  exceeds some  $\epsilon > 0$  behaves like  $\exp(-T \times I(\epsilon))$  where  $I(\epsilon)$ , called large-deviations rate, is the Legendre transform of the topological pressure [13].

When  $T$  is large we have:

$$\mu \left[ \left| \pi_\omega^{(T)}[\mathcal{O}_k] - \mu[\mathcal{O}_k] \right| > \epsilon \right] \simeq \exp\left(\frac{-T \times \epsilon^2}{\sigma(\mathcal{O}_k)}\right) \quad (30)$$

where  $\sigma(\mathcal{O}_k) = \sqrt{\mu[\mathcal{O}_k](1 - \mu[\mathcal{O}_k])}$  is the mean-square deviations of  $\mathcal{O}_k$ .

As a consequence, to obtain the exact average  $\mu[\mathcal{O}_k]$  from our Monte-Carlo algorithm we would need to take the limits:

$$\lim_{T \rightarrow +\infty} \lim_{N_{seed} \rightarrow +\infty} \lim_{N_{flip} \rightarrow +\infty} \frac{\#(\omega_0^{T-1})}{N_{seed}}, \quad (31)$$

in *that order*: they do not commute. A prominent illustration of this point is illustrated in fig. 4.

For notation homogeneity we note from now on  $T - 1 \equiv N_{times}$  for the raster length. When dealing with numerical simulations with a finite number of sample, the goal is to minimize the probability that the error is bigger than a real number  $\epsilon$ , by suitable choice of:

- The raster length:  $T - 1 = N_{times}$ .
- The number of flips:  $N_{flip}$ .
- The number of seed:  $N_{seed}$ .

Let us now establish a few relations between those parameters. First, it is somewhat evident that  $N_{flip}$  must be at least proportional to  $N \times N_{times}$  in order to give a chance to all spikes in the raster to be flipped at least once. This criterion respects the order of limits in (31).

Since  $\mu$  is ergodic one can in principle estimate the average of observables by taking  $N_{seed} = 1$  and taking  $N_{times}$  large. However, the larger  $N_{times}$  the larger  $N_{flip}$  and too big  $N_{times}$  leads to too long simulations. On the opposite, one could generate a large number  $N_{seed}$  of raster with a small  $N_{times}$ . This would have the advantage of reducing  $N_{flip}$  as well. However, the error (30) would then be too large. So, one needs to find a compromise:  $N_{times}$  large enough to have small Gaussian fluctuations (30) and small enough to limit  $N_{flip}$ . Then, by increasing  $N_{seed}$ , one approaches the optimal bound on fluctuations given by (30). Additionally, this provides error bars.

## 4 Numerical tests

In this section, performance in terms of convergence rate and CPU time for increasing values of  $N$  (number of neurons) are discussed. First, we consider potentials (6) where the  $\mathcal{O}_k$ s are observables of the form (1), ("polynomial potentials") and we compare the Monte-Carlo results to those obtained using the transfer matrix and Perron - Frobenius theorem. As discussed in section 3.1 the transfer matrix method becomes rapidly numerically intractable, so that the comparison between Monte-Carlo averages and exact averages cannot be done for large  $N$ . To circumvent this problem, we introduce, in section 4.2 a specific class of potentials for which the analytical computation of the topological pressure as well as observable averages can be analytically done, whatever  $N, R$ . This provides another series of tests.

### 4.1 Polynomial potentials

In this section, we present Monte-Carlo simulations with a potential of the form (6) where the  $\mathcal{O}_k$ s are 100 observables randomly chosen among the  $2^{NR}$  possibilities. More precisely, we select randomly a fraction  $\frac{1}{R}$  of "rates", a fraction  $\frac{1}{R}$  of pairwise terms and so on.

#### 4.1.1 Checking observables average

The figure 2 shows the comparison between the exact values of the observable averages (ordinate) and the estimated Monte-Carlo values (abscissa). The error bars have been computed with  $N_{seed}$  samples. The tests were performed with  $N_{seed} = 20$ ,  $N_{times} = 10000$  and  $N_{flip} = 100000$ . In this example,  $N$  goes from 3 to 8 and  $R = 3$ . For larger values of  $NR$  the numerical computation with the transfer matrix method is not possible any more ( $NR = 24$  corresponds to matrices of size  $16777216 \times 16777216$ ).

#### 4.1.2 Convergence rate

In this section, we show how the Kulback-Leibler divergence varies as a function of  $N_{times}$ . The figure 3 shows the evolution of the Kulback-Leibler divergence between the real distribution and its estimation with the Monte-Carlo method.

We also consider the error

$$error = \max_{k=1\dots K} \left| 1 - \frac{\pi^{(T)}[\mathcal{O}_k]}{\mu[\mathcal{O}_k]} \right|. \quad (32)$$

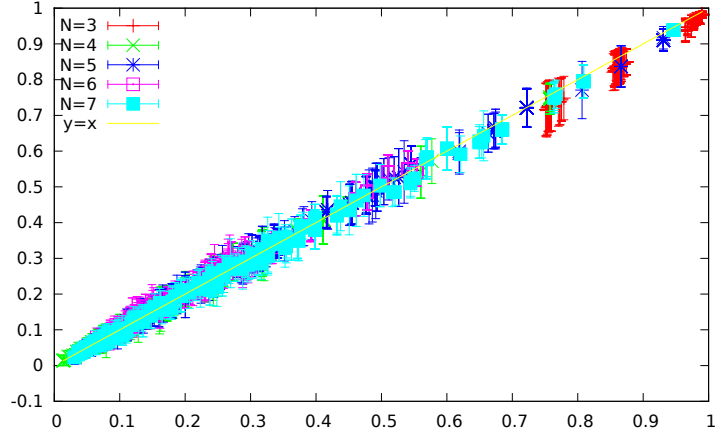


Figure 2: Comparison between the estimated and real values of observable averages.

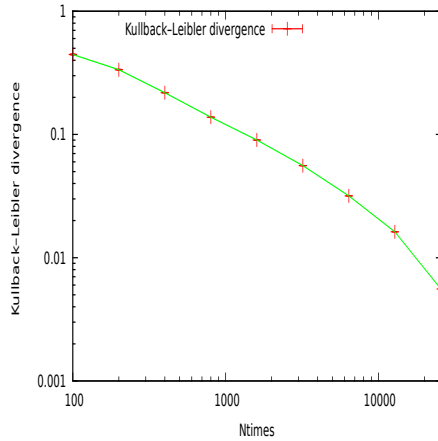


Figure 3: Evolution of Kulback-Leibler divergence 28 as a function  $N_{times}$ .

As developed in section 3.3, this quantity is expected to converge to 0 if  $N_{times} \rightarrow +\infty$  when  $N_{flip}$  grows proportional to  $N \times N_{times}$ . For finite  $N_{times}$ ,  $N_{flip} \rightarrow +\infty$  the error is controlled by the central limit theorem. The probability that the error on the average of observable  $\mathcal{O}_k$  is bigger than  $\epsilon$  (eq. (30)), behaves like  $\exp\left(\frac{-N_{times} \times \epsilon^2}{\sigma(\mathcal{O}_k)}\right)$ .

On the opposite, if  $N_{flip}$  stays constant while  $N_{times}$  grows, the error is expected to first decrease up to a minimum after which it increases. This is because the number of flips is insufficient to reach the equilibrium distribution

of the Monte-Carlo-Hastings Markov chain. This effect is presented in fig. 4. It shows the error (32) as a function of  $N_{times}$  for  $N = 3$  to  $N = 7$  neurons with 3 different  $N_{flip}$  values (1000, 10000, 100000).

Clearly, the number of flips  $N_{flip}$  should be at least more than  $N \times N_{times}$  in order to give a chance to all spikes in the raster to be flipped at least once. A value  $N_{flip} = 10 \times N \times N_{times}$  seems to be enough and computationally reasonable to perform the estimations. With an  $N_{seed} = 20$ , we have results with a reasonable error around the mean values.

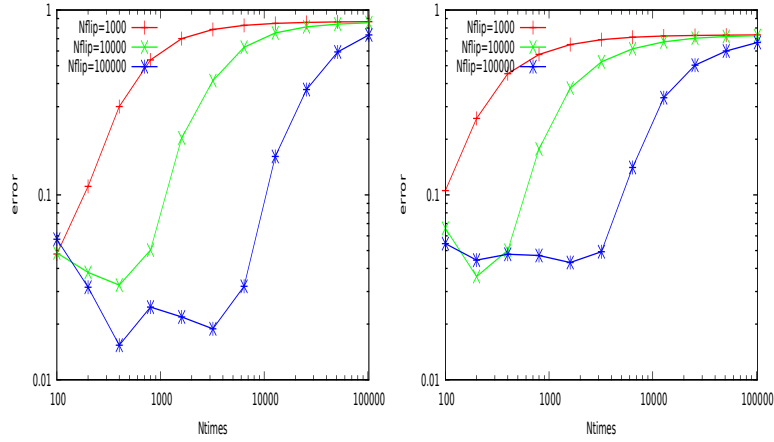


Figure 4: Error as a function of  $N_{times}$ , for several values of  $N_{flip}$  (1000, 10000, 100000). (left)  $N = 3$ ; (right)  $N = 7$ .

### 4.1.3 CPU time

Here we compare the CPU time for a Monte-Carlo simulation and the time for a computation with the transfer matrix. The figure 5 illustrates this. We have plotted the CPU time necessary to obtain the observables average presented in fig. 2, for the Monte-Carlo Average and for the exact average, as a function of  $N$ . The CPU time for Monte-Carlo increases slightly more than linearly while the CPU time for transfer matrix method increases exponentially fast: note that  $R = 3$  here so that  $R \log 2 = 2.08$ , close from the exponential rate found by fit: 2.24.

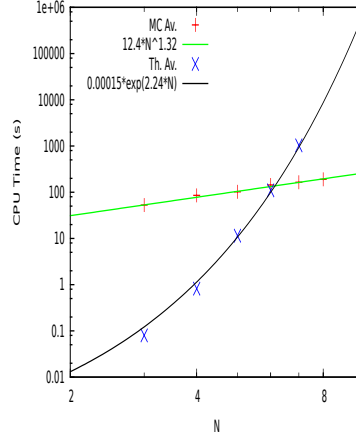


Figure 5: The CPU time necessary to obtain the observable averages presented in fig. 2, for the Monte-Carlo Average (MC Av.) and for the exact average (Th. Av.), as a function of  $N$ . The full lines correspond to fit.

We also plot in fig. 6 the CPU times as a function of  $N_{times}$  with three  $N_{flip}$  values (1000, 10000, 100000), for 3 and 7 neurons. The CPU time increases in a linear fashion with the  $N_{times}$  value. The CPU time also increases linearly with the  $N_{flip}$  value (for the same  $N$  value). The simulations have been done on a computer with the following characteristics: 7 Intel(R) Xeon(R) 3.20GHz processors with a 31.5 Gb RAM.

The figure 7 compares the CPU time increase with  $N_{times}$  for several  $N$  values. It show that the CPU time increases linearly with  $N_{times}$  (as figure 6). However, the slope increases with the number of neurons  $N$ .

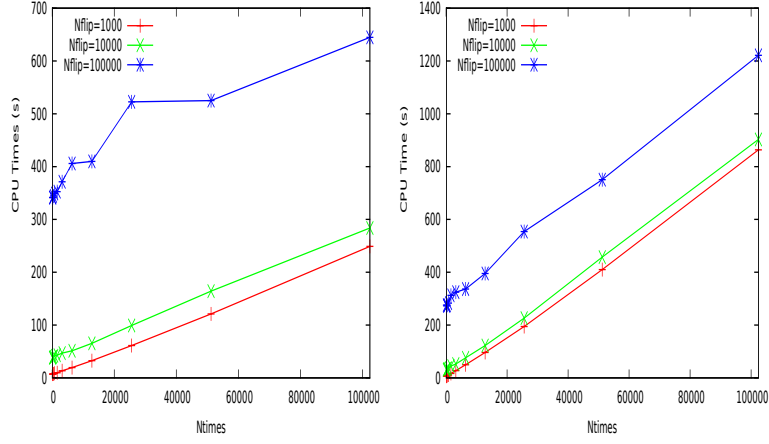


Figure 6: The CPU time ( $T_{cpu}$ ) as a function of  $N_{times}$ .  $T_{cpu}$  increases in a linear fashion with  $N_{time}$  as  $a \times N + b$  where  $a$  is a function of  $N$  and  $b$  is a function of  $N_{flip}$ .

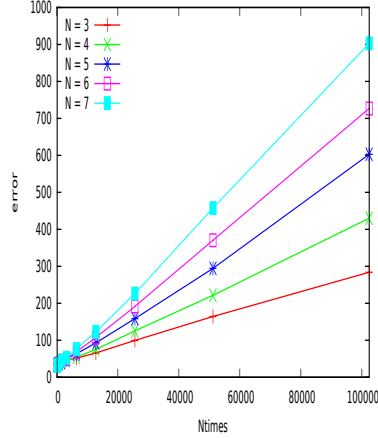


Figure 7: The CPU time ( $T_{cpu}$ ) as a function of  $N_{times}$  for several  $N$  values ( $N_{flip} = 10000$ ).

## 4.2 An analytically solvable example as a benchmark

In this section we consider a specific example of potentials for which the topological pressure is analytically computable, whatever  $N, R$ . As a consequence the average of observables and fluctuations can also be computed. This example is obviously rather specific, but its main interests are to provide a didactic illustration of thermodynamic formalism application as well

as a benchmark for numerical methods.

#### 4.2.1 Analytical setting

We fix the number of neurons  $N$  and the range  $R$  and we choose  $L$  distinct pairs  $(i_l, t_l)$ ,  $l = 1 \dots L$ ,  $i_l \in \{1, \dots, N\}$ ,  $t_l \in \{0, \dots, D-1\}$ . To this set is associated a set of  $K = 2^L$  events  $\mathcal{E}_k = (\omega_{i_1}(t_1), \dots, \omega_{i_L}(t_L))$ ,  $k = 0, \dots, K-1$ . For example, if  $L = 2$ , there are 4 possible events  $\mathcal{E}_0 = (0, 0)$  : neuron  $i_1$  is not firing at time  $t_1$  and neuron  $i_2$  is not firing at time  $t_2$ ;  $\mathcal{E}_1 = (0, 1)$  : neuron  $i_1$  is not firing at time  $t_1$  and neuron  $i_2$  is firing at time  $t_2$ ; and so on. It is convenient to have a label  $k$  corresponding to the binary code of the event.

We define  $K$  observables  $\mathcal{O}_k$  of range  $D$  taking binary values 0, 1. For a block  $\omega_0^D$ ,  $\mathcal{O}_k[\omega_0^D] = 0$  if the event  $\mathcal{E}_k$  is not realized in the bloc  $\omega_0^D$  and is 1 otherwise. In the example above,  $\mathcal{O}_0[\omega_0^D] = 1$  if neuron  $i_1$  is not firing at time  $t_1$  and neuron  $i_2$  is not firing at time  $t_2$  in the block  $\omega_0^D$ . Thus, for  $N = 3$ ,  $R = 4$ ,  $(i_1, t_1) = (1, 0)$ ;  $(i_2, t_2) = (1, 1)$ ,

$$\mathcal{O}_0 \left[ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \right] = \mathcal{O}_0 \left[ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} \right] = 1,$$

while

$$\mathcal{O}_0 \left[ \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \right] = \mathcal{O}_0 \left[ \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix} \right] = 0.$$

We finally define a potential  $\mathcal{H}$  as in (6),  $\mathcal{H}_\beta = \sum_{k=1}^K \beta_k \mathcal{O}_k$ .

For this type of potentials, whatever  $\omega_0^{D-1}$ ,  $\sum_{\omega_0^{D-1}} e^{\mathcal{H}_\beta(\omega_0^D)}$  is a independent of  $\omega(D)$ . As a consequence of the Perron-Frobenius theorem  $s_\beta = \sum_{\omega_0^{D-1}} e^{\mathcal{H}_\beta(\omega_0^D)}$  and therefore:

$$\mathcal{P}(\mathcal{H}_\beta) = (N - L) \log(2) + \log \left[ \sum_{k=1}^K e^{\beta_k} \right].$$

As a consequence, from (24), giving the average of observable  $\mathcal{O}_k$  as the derivative of  $\mathcal{P}$  with respect to  $\beta_k$ :

$$\mu[\mathcal{O}_k] = \frac{e^{\beta_k}}{\sum_{n=1}^K e^{\beta_n}}$$

The fluctuations of observables can also be estimated as well. They are Gaussian with a covariance matrix given by the Hessian of  $\mathcal{P}$  (see eq. (25)) and the central limit theorem.

**Remarks.**

- An important assumption here is that observables do not depend on  $\omega(D)$ . This important simplification as well as the specific form of observables makes the computation of  $\mathcal{P}$  tractable. Note however that, although  $\mathcal{H}$  does not depend on  $\omega(D)$  as well, the *normalized* potential and therefore the conditional probability  $P[\omega(D) | \omega_0^{D-1}]$  depend on  $\omega(D)$  thanks to the normalization factor  $\mathcal{G}$  and its dependence in the right eigenvector  $R$  of the Perron-Frobenius matrix.

#### 4.2.2 Numerical results for large scale networks

Let us now use this example as a benchmark for our Monte-Carlo method. We have considered a case with range  $R = 4$  and  $L = 6$  pairs, corresponding to  $2^6 = 64$  terms in the potential. We have analyzed the convergence rate as a function of  $N$  the number of neurons and  $N_{times}$ , the time length of the Monte-Carlo raster. The number of flips,  $N_{flip}$  is fixed to  $10 \times N \times N_{times}$ , so that, on average, each spike of the Monte-Carlo raster is flipped 10 times in one trial.

In fig. 8 (left) we have shown the relative error (eq. 32) as a function of  $N_{times}$  for several values of  $N$ . The empirical average  $\pi^{(T)}[\mathcal{O}_k]$  is computed on 10 Monte-Carlo rasters. That's why we don't write the index  $\omega$  in the empirical probability. We stopped the simulation when the error is lower than 5%. As expected from the Central Limit Theorem (CLT), the error decreases as a power law  $C N_{times}^{-\frac{1}{2}}$  where the constant  $C$  has been obtained by fit.

In fig. 8 (right) we have drawn the CPU time as a function of  $N_{times}$  for several values of  $N$ . It increases linearly with  $N_{times}$ , with a coefficient depending on  $N$ . The simulation is relatively fast: it takes 2h30 for  $N = 60$ ,  $N_{times} = 8000$  (with 10 Monte-Carlo trials) on a 7 processors machine (each of these processors has the following specifications: Intel(R) Xeon(R) CPU 2.27GHz, 1.55 MB of each memory size) with a 17.72 GB RAM.

## 5 Discussion and perspectives

In this paper, we have shown how maximum entropy models can be extended to analyse the spatio-temporal dynamics of the neural activity. This raises specific issues, mainly related to the fact that the normalization of the Gibbs potential depends on the past activity. We have shown that transfer matrices results allow to handle this problem properly, providing additionally crucial information on statistics (especially average of observables and fluctuations of empirical averages). The challenge is then to be able to fit these models



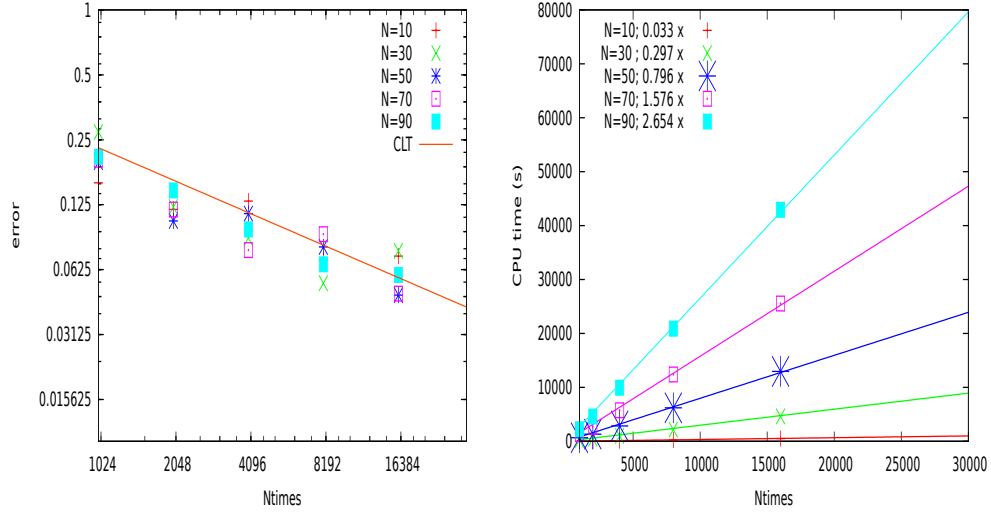


Figure 8: (Left). Relative error as a function of  $N_{times}$  for several values of  $N$ . (CLT) indicates the decay expected from the central limit theorem. As expected from the Central Limit Theorem, the error decreases as power law  $N_{times}^{-\frac{1}{2}}$ . (Right). CPU time as a function of  $N_{times}$  for several values of  $N$ . It increases linearly as  $CPU = a \times N_{times}$ . In the legend, next to the value of  $N$  we indicate the value of  $a$ .

to the recordings. A major step in the fitting process is to compute the observables generated by the model for a given set of parameters. We have proposed a first method, based on the transfer matrix. It gives exact results, but can only be applied to small subsets of neurons. We have then designed a Monte-Carlo approach that overcomes this issue, confirmed by several tests.

In fact, matching Gibbs averages of observables is only a first, although crucial, step towards spike train analysis of neuronal activity. The next step consists of fitting the parameters of a model from an experimental raster. Basically, this corresponds to minimizing the Kullback-Leibler divergence (28) between the model and the empirical measure. We have reviewed some possible techniques in the section 2.3.4. The application of our method to fit Gibbs distributions on large-scale retina recordings will be considered in a forthcoming paper.

As a final issue we would now like to discuss shortcomings of Maximum entropy Models. Although initially proposed as powerful methods for neuroscience applications, many future reports have cast doubt on how useful (spatial pairwise) Maximum Entropy models were. These criticism include

the role of common input [35], the role of higher order correlations [67, 42], scaling properties [52] and non stationarity [67]. As a matter of fact, the models presented in the section 2.3.5 are clear and efficient alternatives to spatial pairwise maximum entropy models. Let us now comment these shortcomings.

First, as we have developed, the maximum entropy approach is definitely not limited to spatial pairwise interactions. Especially, the role of higher order interactions beyond pairwise equal time ones, provides a clear motivation for including longer temporal history in statistical models of neural data.

This raises however the question of how one chooses the potential. As exposed in section 2.2.1 the possible number of constraints is simply overwhelming and one has to make choices to reduce their number. These choices can be based on ad hoc assumptions (e.g. rates or instantaneous pairwise correlations are essential in neuronal coding) or on empirical constraints (type of cells, spatial localisation). However, this combinatorial complexity is clearly a source of troubles and questions about the real efficiency of the maximum entropy problem. This problem is particularly salient when dealing with temporal interactions of increasing memory: even the number of possible pairwise interactions might be too large to fit all of them on finite size recordings. Additionally, using a too complex potential increases the number of parameters necessary to fit the data beyond what the number of available samples allows.

One solution is to try to infer the form of the potential from the data set. Important theorems in ergodic theory [50] as well as in Variable Length Markov Chains Estimations can be used [10]. This will be developed in a separate paper.

It is however quite restrictive to stick to potentials expressed as linear combinations of observables, like (6). This form has its roots in thermodynamics and statistical physics, but is far from being the most general form. Nonlinear potentials such as (27) (GLM) can be considered as well. Although such potentials can be expressed in the form (6) from the Hammersley-Clifford theorem [23], this representation induces a huge redundancy in the coefficients  $\beta_k$ . Examples are known of non linear potentials with relatively small number of parameters  $\lambda_l$ , which, expressed in the form (6), give rise to  $2^{NR}$  parameters  $\beta_k$ s, all of them being functions of the  $\lambda_l$ s: see [11, 12]. Such potentials constitute relevant alternatives to (6) where the formalism described here fully applies.

More generally, alternatives to maximum entropy models consider different models trying to mimic the origin of the observed correlations. This is the case of the model proposed by [36] when common inputs are added to account for the instantaneous correlations and the GLM model where the

numbers of parameters is only  $N^2$ . However, note that in all these cases, the models constrain the correlations to be in a specific form, and might not be a good description of the activity either. Testing these models on data is the only way to distinguish the most relevant ones. Note that the discrepancy between model and data will probably be more and more obvious with a larger set of neurons. The validity of a model will also depend on size of the recorded population.

Another, even deeper, question is the translation-invariance assumption intrinsic to the maximum entropy principle. When dealing e.g. to transient responses to temporary stimuli this assumption is clearly highly controversial. Note however that although the maximum entropy principle does not extend to non translation-invariant statistics, the concept of Gibbs distribution extend to that case [22]. Here, Gibbs distributions are constructed via transition probabilities, possibly with an infinite memory. Examples of applications to neuronal networks can be found in [11, 12]. However, the application of this concept to analyzing real data, especially the problem of parameters estimation, remains to our knowledge an open challenge.

**Acknowledgments** We are grateful to G. Tkacik, T. Mora, S. Kraria, T. Viéville, F. Hebert for helpful advices and help. This work was supported by the INRIA, ERC-NERVI number 227747, KEOPS ANR-CONICYT and European Union Project # FP7-269921 (BrainScales) projects to B.C and H.N. and ANR OPTIMA to O.M. Finally, we would like to thank the reviewers for helpful comments and remarks.

## References

- [1] M. Abeles. Local Cortical Circuits: An Electrophysiological study. Springer, Berlin, 1982.
- [2] H. Ackley, E. Hinton, and J. Sejnowski. A learning algorithm for boltzmann machines. Cognitive Science, pages 147–169, 1985.
- [3] H. Ackley, E. Hinton, and J. Sejnowski. a mean field theory learning algorithm for neural network. Complex systems, pages 995–1019, 1987.
- [4] Yashar Ahmadian, Jonathan W. Pillow, and Liam Paninski. Efficient Markov Chain Monte Carlo Methods for Decoding Neural Spike Trains. Neural Computation, 23(1):46–96, January 2011.
- [5] C. Beck and F. Schloegl. Thermodynamics of Chaotic Systems: An Introduction. Cambridge University Press, Cambridge, 1995.

- [6] D. R. Brillinger. Maximum likelihood analysis of spike trains of interacting nerve cells. Biol Cybern, 59(3):189–200, 1988.
- [7] D. R. Brillinger. Nerve Cell Spike Train Data Analysis - a Progression of Technique. J Amer Statist Assn, 87(418):260–271, 1992.
- [8] Tamara Broderick, Miroslav Dudík, Gasper Tkacik, Robert E. Schapire, and William Bialek. Faster solutions of the inverse pairwise ising problem. Submitted (see <http://arxiv.org/abs/0712.2437>), 2007.
- [9] E. N. Brown, R. Barbieri, U. T. Eden, and L. M. Frank. Likelihood methods for neural spike train data analysis. Computational Neuroscience: A Comprehensive Approach, 2003.
- [10] Peter Buehlmann and Abraham J. Wyner. Variable length markov chains. The Annals of Statistics, 27(2), 1999.
- [11] Bruno Cessac. A discrete time neural network model with spiking neurons ii. dynamics with noise. J. Math. Biol., 62:863–900, 2011.
- [12] Bruno Cessac. Statistics of spike trains in conductance-based neural networks: Rigorous results. Journal of Mathematical Neuroscience, 1(8), 2011.
- [13] J.R. Chazottes and G. Keller. Pressure and equilibrium states in ergodic theory. Israel Journal of Mathematics, 131(1), 2008.
- [14] E. J. Chichilnisky. A simple white noise analysis of neuronal light responses. Network: Comput. Neural Syst., 12:199–213, 2001.
- [15] Simona Cocco, Stanislas Leibler, and Rémi Monasson. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. PNAS, 106(33):14058–14062, 2009.
- [16] Rodrigo Cofré and Bruno Cessac. Dynamics and spike trains statistics in conductance-based integrate-and-fire neural networks with chemical and electric synapses. Chaos, Solitons and Fractals, submitted, 2012. submitted.
- [17] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai. Ergodic Theory. Springer, Berlin, Heidelberg, New York, 1982.
- [18] Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In Proceedings of the 17th Annual Conference on Computational Learning Theory, 2004.
- [19] Elad Ganmor, Ronen Segev, and Elad Schneidman. The architecture of functional interaction networks in the retina. The journal of neuroscience, 31(8):3044–3054, 2011.

- [20] Elad Ganmor, Ronen Segev, and Elad Schneidman. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. PNAS, 108(23):9679–9684, 2011.
- [21] F. R. Gantmacher. the theory of matrices. AMS Chelsea Publishing, Providence, RI, 1998.
- [22] Hans-Otto Georgii. Gibbs measures and phase transitions. De Gruyter Studies in Mathematics:9. Berlin; New York, 1988.
- [23] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. unpublished, 1971.
- [24] K. D. Harris, D. A. Henze, H. Hirase, X. Leinekugel, G. Dragoi, A. Czurko, , and G. Buzsaki. Spike train dynamics predicts theta-related phase precession in hippocampal pyramidal cells. Nature, 417(6890):738–741, 2002.
- [25] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. Biometrika, 57:97–109, 1970.
- [26] S Higuchi and M Mezard. Susceptibility propagation for constraint satisfaction problems. CoRR, pages –1–1, 2009.
- [27] Y. Ikegaya, G. Aaron, R. Cossart, D. Aronov, I. Lampl, D. Ferster, , and R. Yuste. Synfire chains and cortical songs: Temporal modules of cortical activity. Science, 304(5670):559–564, 2004.
- [28] E.T. Jaynes. Information theory and statistical mechanics. Phys. Rev., 106:620, 1957.
- [29] H.J. Kappen and F.B. Rodriguez. Boltzmann machine learning using mean field theory and linear response correction. In Advances in Neural Information Processing Systems, pages 280–286. The MIT Press, 1998.
- [30] G. Keller. Equilibrium States in Ergodic Theory. Cambridge University Press, 1998.
- [31] T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, , and A. Arieli. Spontaneously emerging cortical representations of visual attributes. Nature, 425(6961):954–956, 2003.
- [32] I. Lampl, I. Reichova, and D. Ferster. Synchronous membrane potential fluctuations in neurons of the cat visual cortex. Neuron, 22(2):361–374, 1999.
- [33] K. Louie and M. A. Wilson. Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. Neuron, 29(1):145–156, 2001.

- [34] A. Luczak, P. Barthó, and K. D Harris. Spontaneous events outline the realm of possible sensory responses in neocortical populations. Neuron, 62(3):413–425, 2009.
- [35] Jakob H. Macke, John P. Cunningham, Krishna V. Shenoy, Lars BÅ $\frac{1}{4}$ sing, Byron M. Yu, and Maneesh Sahani. Empirical models of spiking in neural populations, 2010.
- [36] J.H Macke, L Busing, J.P Cunningham, B.M Yu, K.V Shenoy, and M Sahani. Empirical models of spiking in neural populations. NIPS, 2001.
- [37] O. Marre, S. El Boustani, Y. Frégnac, and A. Destexhe. Prediction of spatiotemporal patterns of neural activity from pairwise correlations. Phys. rev. Let., 102:138101, 2009.
- [38] P. McCullagh and J. A. Nelder. Generalized linear models (Second edition). London: Chapman & Hall, 1989.
- [39] M. Mézard and T. Mora. Constraint satisfaction problems and neural networks: a statistical physics perspective. J. Physiol. Paris, 103:107–113, 2009.
- [40] A. Mokeichev, M. Okun, O. Barak, Y. Katz, O. Ben-Shahar, and I. Lampl. Stochastic emergence of repeating cortical motifs in spontaneous membrane potential fluctuations in vivo. Neuron, 53(3):413–425, 2007.
- [41] Sheila H. Nirenberg and Jonathan D. Victor. Analyzing the activity of large populations of neurons: how tractable is the problem? Current Opinion in Neurobiology, 17(4):397–400, 2007.
- [42] Ifije E. Ohiorhenuan, Ferenc Mechler, Keith P. Purpura, Anita M. Schmid, Qin Hu, and Jonathan D. Victor. Sparse coding and high-order correlations in fine-scale cortical networks. Nature, 466(7306):617–621, 2010.
- [43] M. W. Oram, M. C. Wiener, R. Lestienne, and B. J. Richmond. Stochastic nature of precisely timed spike patterns in visual system neuronal responses. J Neurophysiol, 81(6):3021–3033, 1999.
- [44] L. Paninski, M. Fellows, S. Shoham, N. Hatsopoulos, and J. Donoghue. Superlinear population encoding of dynamic hand trajectory in primary motor cortex. J. Neurosci., 24:8551–8561, 2004.
- [45] Liam Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. Network: Comput. Neural Syst., 15(04):243–262, November 2004.
- [46] W. Parry and M. Pollicott. Zeta functions and the periodic orbit structure of hyperbolic dynamics, volume 187–188. Asterisque, 1990.

- [47] J W Pillow, J Shlens, L Paninski, A Sher, A M Litke, E J Chichilnisky, and E P Simoncelli. Spatio-temporal correlations and visual signaling in a complete neuronal population. Nature, 454(7206):995–999, Aug 2008.
- [48] Jonathan W. Pillow, Yashar Ahmadian, and Liam Paninski. Model-based decoding, information estimation, and change-point detection techniques for multineuron spike trains. Neural Comput., 23(1):1–45, 2011.
- [49] J.W. Pillow, L. Paninski, V.J. Uzzell, E.P. Simoncelli, and E.J. Chichilnisky. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. Journal of Neuroscience, 25(47):11003–11013, 2005.
- [50] Mark Pollicott and Howard Weiss. Free energy as a dynamical invariant (or can you hear the shape of a potential?). Communications in Mathematical Physics, 240:457–482, 2003.
- [51] J. L. Puchalla, E. Schneidman, R. A. Harris, and M. J. Berry. Redundancy in the population code of the retina. Neuron, 46(3):493–504, 2005.
- [52] Y. Roudi, S. Nirenberg, and P.E. Latham. Pairwise maximum entropy models for studying large biological systems: when they can work and when they can’t. PLOS Computational Biology, 5(5), 2009.
- [53] Yasser Roudi and John Hertz. Mean field theory for non-equilibrium network reconstruction. Phys. Rev. Lett., 106(048702), 2011.
- [54] Yasser Roudi, Joanna Tyrcha, and John A Hertz. Ising model for neural data: Model quality and approximate methods for extracting functional connectivity. Physical Review E, page 051915, 2009.
- [55] D. Ruelle. Thermodynamic formalism. Addison-Wesley, Reading, Massachusetts, 1978.
- [56] E. Schneidman, M.J. Berry, R. Segev, and W. Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. Nature, 440(7087):1007–1012, 2006.
- [57] E. Seneta. Non-negative Matrices and Markov Chains. Springer, 2006.
- [58] V Sessak and R Monasson. Small-correlation expansions for the inverse ising problem. Journal of Physics A, 42(055001), 2009.
- [59] M. Shadlen and W. Newsome. The variable discharge of cortical neurons: Implications for connectivity. J. Neurosci, 18:3870–3896, 1998.
- [60] Krishna V. Shenoy, Matthew T. Kaufman, Maneesh Sahani, and Mark M. Churchland. dynamical systems view of motor preparation: Implications for neural prosthetic system design. Progress in Brain Research: Enhancing Performance for Action and Perception, 192, 2011.

- [61] D. Sherrington and S. Kirkpatrick. Solvable model of a spin-glass. Physical Review Letters, 35(26):1792+, December 1975.
- [62] J. Shlens, G.D. Field, J.L. Gauthier, M.I. Grivich, D. Petrusca, A. Sher, A.M. Litke, and E.J. Chichilnisky. The structure of multi-neuron firing patterns in primate retina. Journal of Neuroscience, 26(32):8254, 2006.
- [63] Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Martin Greschner, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of large-scale synchronized firing in primate retina. The Journal of Neuroscience, 29(15):5022–5031, April 2009.
- [64] E. P. Simoncelli, J. P. Paninski, J. Pillow, and O. Schwartz. Characterization of Neural Responses with Stochastic Stimuli. The cognitive neurosciences, 2004.
- [65] W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. Annual Review of Neuroscience, 18(1):555–586, 1995.
- [66] William R. Softky and Christof Koch. The highly irregular firing of cortical cells is inconsistent with temporal integration of random epsps. Journal of Neuroscience, 13:334–350, 1993.
- [67] B. Staude, S. Grun, and S. Rotter. Higher-order correlations in non-stationary parallel spike trains: statistical modeling and inference. Frontiers in computational neuroscience, 4, 2010.
- [68] S.P. Strong, R. Koberle, R.R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. Phys. Rev. Let., 80(1):197–200, 1998.
- [69] T. Tanaka. A theory of mean field approximation. ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 48:351–360, 1998.
- [70] Aonan Tang, David Jackson, Jon Hobbs, Wei Chen, Jodi L. Smith, Hema Patel, Anita Prieto, Dumitru Petrusca, Matthew I. Grivich, Alexander Sher, Pawel Hottowy, Wladyslaw Dabrowski, Alan M. Litke, and John M. Beggs. A maximum entropy model applied to spatial and temporal correlations from cortical networks *In Vitro*. The Journal of Neuroscience, 28(2):505–518, January 2008.
- [71] F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network, 12(3):289–316, January 2001.
- [72] D. J. Thouless, P. W. Anderson, and R. G. Palmer. Solution of solvable model of a spin glass. Phil. Mag., 35:593–601, 1977.



- [73] Gasper Tkacik, Jason S. Prentice, Vijay Balasubramanian, and Elad Schneidman. Optimal population coding by noisy spiking neurons. PNAS, 107(32):14419–14424, August 2010.
- [74] Gasper Tkacik, Elad Schneidman, Michael J. Berry II, and William Bialek. Spin glass models for a network of real neurons. arXiv: 0912.5409v1, 2009.
- [75] Wilson Truccolo, Uri T. Eden, Matthew R. Fellows, John P. Donoghue, and Emery N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. J Neurophysiol, 93:1074–1089, 2005.
- [76] M. Tsodyks, T. Kenet, A. Grinvald, , and A. Arieli. Linking spontaneous activity of single cortical neurons and the underlying functional architecture. Science, 286(5446):1943–1946, 1999.
- [77] E. Vaadia, I. Haalman, M. Abeles, H. Bergman, Y. Prut, H. Slovin, and A. Aertsen. Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. Nature, 373(6514):515–518, 1995.
- [78] Juan Carlos Vasquez, Olivier Marre, Adrian G Palacios, Michael J Berry, and Bruno Cessac. Gibbs distribution analysis of temporal correlation structure on multicell spike trains from retina ganglion cells. J. Physiol. Paris, 2012. in press.
- [79] Max Welling and Yee Whye Teh. Approximate inference in boltzmann machines. Artificial Intelligence, 143(1):19 – 50, 2003.
- [80] Shan Yu, Debin Huang, Wolf Singer, and Danko Nikolic. A small world of neuronal synchrony. Cereb. Cortex, 2008.